Scale-banking for Patient Reported Outcome Measures (PROMs) measuring functioning in rheumatoid arthritis : a daily activities metric

Prodinger, B, Coenen, M, Hammond, A, Küçükdeveci, AA and Tennant, A

http://dx.doi.org/10.1002/ACR.24503

Accepted article: https:// doi:10.1002/ACR.24503. Arthritis Care and Research. First published online 5 November 2020.

Scale-Banking for Patient Reported Outcome Measures (PROMs) Measuring Functioning in Rheumatoid Arthritis: A Daily Activities Metric

Prof. Birgit Prodinger, PhD, MSc, MSc (Corresponding Author)
Faculty of Applied Health and Social Sciences, University of Applied Sciences Rosenheim
Hochschulstr. 1, 83024 Rosenheim, Germany
E-mail: birgit.prodinger@th-rosenheim.de
Phone: +49 8031 805 2263
Fax: +49 8031 805 2786
Swiss Paraplegic Research, Guido A. Zäch Str. 4, 6207 Nottwil, Switzerland
ICF Research Branch, Guido A. Zäch Str. 4, 6207 Nottwil, Switzerland

Dr. Michaela Coenen, MPH
Institute for Medical Information Processing, Biometry, and Epidemiology – IBE
Chair of Public Health and Health Services Research, LMU Munich
Pettenkofer School of Public Health, Munich, Germany
Marchioninistr. 17, 81377 Munich, Germany
ICF Research Branch, Guido A. Zäch Str. 4, 6207 Nottwil, Switzerland

Prof. Alison Hammond, PhD
Centre for Health Sciences Research, University of Salford,
Allerton L701, Salford, M6 6PU, UK

Prof. Ayşe A. Küçükdeveci, MD
Faculty of Medicine, Department of Physical Medicine and Rehabilitation, Ankara University
Samanpazari, 06100 Ankara, Turkey

Prof. Emeritus Alan Tennant, PhD
Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds
Leeds, LS2 9JT, United Kingdom
Swiss Paraplegic Research, Guido A. Zäch Str. 4, 6207 Nottwil, Switzerland
ICF Research Branch, Guido A. Zäch Str. 4, 6207 Nottwil, Switzerland

Competing interests: The authors declare that they have no competing interests.

Funding: This study was funded by a EULAR Health Professional Grant.

Acknowledgements: The authors would like to express their gratitude to all participants in this study.

Word Count: 3805 words

Abstract

Objective: Functioning is an important outcome for rheumatoid arthritis (RA) management. Heterogeneity of respective patient-reported outcome measures (PROMs) challenges direct comparisons between their results. This study aimed to standardize reporting of such PROMs measuring functioning in RA to facilitate comparability.

Methods: Common Item Non-Equivalent Groups Design (NEAT) with the Health Assessment Questionnaire (HAQ) as a common scale across data sets from various countries (incl. UK, Turkey and Germany) to establish a common metric. Other PROMs included are the Physical Function items of the Multidimensional Health Assessment Questionnaire (MDHAQ), Disabilities of Arm, Shoulder and Hand (DASH), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), World Health Organization Disability Assessment Schedule Version 2.0 (WHODAS 2.0), and four short forms (20, 10, 6, and 4 physical function items) from the Patient-Reported Outcomes Measurement Information System (PROMIS). As the HAQ includes mobility, self-care and domestic life items, this study focuses on these three domains. PROMs were described using Standard Error of Measurement (SEM) and Smallest Detectable Difference (SDD). Rasch Measurement model was used to create the common metric.

Results: Range of SEM is 0.2 (MDHAQ) to 7.4 (SF36-PF). SDD revealed a range from 9.7 % (WOMAC-RAT) to 33.5 % (WHODAS-PF). PROMs co-calibration revealed fit to the Rasch measurement model. A transformation table was developed to allow exchange between PROMs scores.

Discussion: Scores between the Daily Activity PROMs commonly used in RA can now be compared. Factors such as SEM and SDD help determine choice of PROM in clinical practice and research.

Keywords

HAQ
Multidimensional HAQ
DASH
WOMAC
WHODAS 2.0
PROMIS-SF
Rasch measurement model
International Classification of Functioning, Disability and Health
Common metric
Scale banking

Significance and Innovations

- The number and heterogeneity of patient-reported outcome measures (PROMs) used in clinical research and practice in rheumatoid arthritis (RA) makes it difficult to directly compare the results of these PROMs from different settings or studies.
- This study enables direct comparability of commonly used PROMs to assess activities of daily living by means of an interval-scaled Daily Activities Metric.
- The PROMs included in this study all measure a similar range on the Daily Activities Metric, thus other factors, such as the Smallest Detectable Difference (SDD), are suggested to be used to differentiate between PROMs.
- Differences in SDD occurred, whereby the Health Assessment Questionnaire (HAQ) is of particular concern, indicating that it is less than optimal for detecting a difference compared to other PROMs.

Studies of the lived experience of people with Rheumatoid Arthritis (RA) show that most facets of life can be affected by the health condition (1, 2), and thus, are important outcomes to measure in evaluating and monitoring the health condition and related interventions: Body structures and functions can be impaired, activities in daily life limited, as well as social, community, and civic life being restricted. Therefore, a comprehensive understanding of health, as reflected in a bio-psycho-social perspective, is foundational for measuring outcomes in clinical trials, epidemiological studies or the routine monitoring of the patients' progress (3, 4). "Outcome" refers here to any indicator (variable) to detect changes in health status or quality of life. Clinical and researchers use a wide range of outcomes, from inflammatory markers and joint counts through to job retention and quality of life (5-8). Many such outcomes use questionnaires to measure patients' perceptions of the condition's impact on their health and lives. Such Patient Reported Outcome Measures (PROMs) have been used in RA for over 35 years (9). In the context of this study, a PROM is defined as any patient- (or proxy) completed questionnaire in which a set of items is summated to give a total score, a series of 'domain' scores, or both. "Domain" refers to any meaningful aggregation of categories as defined by the International Classification of Functioning, Disability and Health (ICF) (10). ICF categories (e.g. d450 Walking) are the unit of the classification and are hierarchically ordered into chapters (e.g. d4 Mobility) and components (e.g. d Activities & Participation). The components and their interactions reflect a bio-psycho-social model of health and disability in RA (11, 12).

The use of PROMs in rheumatology is ubiquitous. For example, a recent European League Against Rheumatism (EULAR) PROM Program project found, from 2000-2016,78 different PROMs were used to measure outcomes in Osteoarthritis (OA) studies (13). Often several different PROMs can be used to measure the same domain, such as pain, fatigue, mobility or self-care. This heterogeneity makes it difficult to directly compare the results of PROMs from

different studies. Furthermore, data derived from PROMs are often ordinal-scaled, limiting their usefulness in monitoring change over time (14). The lack of comparable and interval-scaled information collected from PROMs measuring the same construct, restricts using data for secondary clinical purposes, such as quality audits and benchmarking, as well as for research purposes, including meta-analyses. However, international standards for eHealth stress the need for information systems based on international health classifications, including the ICF, to ensure that health information is available in a consistent and comparable manner for effective use in decision-making (15). Therefore, the objective of this study was to standardize reporting of commonly used PROMs in RA to facilitate their comparability.

Patients and Methods

Conceptual and score equivalence are foundational to establishing comparability of existing PROMs (16). For conceptual equivalence, we relied on previously linked items from selected PROMs to the ICF (www.icf-research-branch.org). PROMs linked to the same ICF domains are assumed to be comparable from a content perspective and thus could be included in the psychometric analyses to establish score equivalence. The Rasch measurement model was applied, with total PROMs' scores equated directly, to establish score equivalence, rather than ratings of single items within each PROM (17).

Outcome Measures

The 10 most commonly used PROMs in the last 10 years (2006-2016) in RA research were identified based on the preliminary results of the second part of the above mentioned EULAR project focusing on PROMs used in RA. Of those, six include items that were linked to the ICF component Activities and Participation. The remaining four were the EuroQol (not a summated scale and with mixed domain content; 18), Hospital Anxiety and Depression scale (HADS; 19), the Short Form-6 Dimensions (SF-6D), and the Rheumatoid Arthritis Quality of

Life scale (RAQoL; 20). The six chosen included the Medical Outcomes Study 36-item Short Form Health Survey (SF-36) (21); the Health Assessment Questionnaire (HAQ) (9); the Disabilities of Arm, Shoulder, and Hand (DASH) (22); the Multidimensional Health Assessment Questionnaire (MDHAQ) (23); the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) (24); and the World Health Organization Disability Assessment Schedule Version 2.0 (WHODAS 2.0) (25). Other generic PROMs allowing comparability across conditions were included, that is, relevant sub-scales from the Patient-Reported Outcomes Measurement Information System (PROMIS) (26), as it is a recommended PROM for functional status assessment (27).

As the HAQ is the most commonly used PROM in RA and covers mainly activities related to mobility, self-care and domestic life, this study focused on these three ICF Activities and Participation domains. Amongst the selected PROMs, only (sub-)scales that mapped on to the d4 Mobility, d5 Self-Care and d6 Domestic Life domains were chosen. Since items within each included PROM are generally consistent with undertaking tasks associated with Activities of Daily Living, the resulting interval-scaled common metric was referred to as the Daily Activities Metric.

The Health Assessment Questionnaire (HAQ) (9) consists of 20 items assessing difficulties in performing activities of daily living on a scale of 0="Without any difficulty" to 3="Unable to do." These items are grouped into eight domains. To create a total score, the highest item scores from each domain are added and then divided by eight with higher scores indicating more difficulties. In this study, the HAQ was scored without the score adjustment for assistive devices and help, because the other included PROMs reflect a performance perspective, whereas adjusting HAQ scores attempts a capacity perspective, i.e. trying to

ascertain what level of problem the individual would have had without using assistive devices or help.

The Medical Outcomes Study Short Form 36-Item (SF-36; Version 2) (21) comprises eight health domains whereby only the physical functioning (SF36-PF) domain was relevant for this study. The SF36-PF consists of 10 items related to activities of daily living, each rated on a scale from 1="Limited a lot" to 3="Not limited at all". The total score is created by summing the responses to each item and transforming it to a 0-100 scale, with lower scores indicating worse function.

The Disabilities of the Arm, Shoulder, and Hand (DASH) (22) contains 30 items related to physical function and symptoms. Only the 23 items related to physical function were included and rated on a scale from 0="No difficulty" to 5="Extreme difficulty". The mean of the items is transformed into a scale from 0 to 100 for the total score (((sum of n responses-1)/n))*25), with higher scores indicating worse function.

The Multidimensional Health Assessment Questionnaire (MDHAQ) (23) consists of 10 items: the eight MHAQ items plus walking three kilometers and participating in recreational activities. The total score is the sum of the items divided by the total number of items answered (at least nine out of the 10 are required). The value is rounded to the first decimal, with higher scores indicating worse function.

The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) (24) consists of three sub-scales (pain, stiffness, and physical function). Only the physical function sub-scale (WOMAC-PF) which includes 17 items was included. Two forms of the WOMAC-PF are available; one with a Numeric Rating scale scored 0-10 (WOMAC-NRS), the other

with a rating scale scored 0 to 4, whereby 0 always indicates no difficulty and the higher score extreme difficulty (WOMAC-RAT). Since both forms are used in practice, we included both. A total score for each sub-scale is created by summing up the respective items, with higher scores indicating worse function.

The World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) (25) is a generic health and disability instrument with six domains. Three domains (Mobility, Self-care, and Life Activities) equated to the ICF Chapters d4 Mobility, d5 Self Care and d6 Domestic Life and thus were relevant for this study (WHODAS-PF). Items are scored on a scale from 0= "No difficulty" to 4= "Extreme difficulty/cannot do." A total score for each domain is created by summing up its items' responses, with higher scores indicating worse function.

The Patient-Reported Outcomes Measurement Information System (PROMIS) (26) is a set of measures of physical, mental, and social health. In this study, we included the four physical function short-forms (PF-20, PF-10, PF-6, and PF-4 items) of the PROMIS. Items are rated on a scale from 1= "Cannot do" to 5= "Without any difficulty". A total score for each short form is created by summing up the responses to each item, with lower scores indicating worse function.

In total, we included 11 PROMs including four forms of the PROMIS, two forms of the HAQ and two forms of the WOMAC. All PROMs were collected using the validated language version in the participating countries.

Data collection

We adopted a twofold strategy: First, we considered data sets in which data of the identified PROMs was already collected previously and applied for data collected in the process of developing and validating the ICF Core Set for RA at LMU Munich which coordinated the ICF Core Set development process relying on an international network. More specifically, we used the data from Lithuania, Serbia, Hungary and the Netherlands, grouped it together under an "other Europe" label. Participants were diagnosed with RA according to the study criteria of the primary studies.

Second, to ensure that all PROMs, or at least one version of each PROM, was well populated in English, German and Turkish, we collected additional data in people with RA at Ankara University, University of Salford, and LMU Munich (Figure 1). All relevant documents were prepared in a generic form and then adopted to local regulations by the local research teams to ensure that data collection followed the respective regulations in place at the time. Data collection took place between Spring 2017 and 2018 through the outpatient clinic or established patient networks at each site. Ethical approval was obtained from the appropriate Research Ethics Committees at each site and each participant gave his/her informed written consent to participate in this study.

[Figure 1]

Data analysis

Analysis was embedded within a common-item, non-equivalent person design (NEAT) with the HAQ being the common PROM across all datasets. NEAT implies that the same items were administered in different groups, but not all persons have administered all items. This design allows bringing together different data sets containing different PROMs, yet have at least one item set common across all sets – the HAQ in the present study. Descriptive

statistics were used to describe PROMs scores for each country; the Kruskal-Wallis test to determine any differences in the ordinal PROM scores across countries. In addition, the Standard Error of Measurement (SEM; SD* √(1-alpha)) and Smallest Detectable Difference (SDD; SEM*1.96*√2) were calculated on the raw scale scores to gain information about the level of precision of the scale. The SDD was also presented as a percent of the full operational range of the PROM (i.e. its total raw score range). Cronbach's alpha is reported as an indicator of internal reliability of each scale.

To co-calibrate the scales onto a common reference metric (an interval-scaled metric with three or more scales), the Partial Credit parameterisation of the Rasch Measurement model was used within the RUMM2030 software (29, 30). The analytical test-equating approach adopted in this study is recent, involving just the total scores of the scales to represent items within the Daily Activities Metric (17). This has the advantage of absorbing any local item dependency that exists within each scale. Thus, the scales intended to measure the Daily Activities domain were calibrated onto the reference metric and their fit to the Rasch model tested as a set of items, that is, each PROM represented an item.

Due to the incomplete nature of the data matrix (not all PROMs were collected in each setting), fit to the model was tested by pairwise PROM fit, with the HAQ always being present. Such a pairwise test of fit makes available a robust Conditional Test of Fit (CTF) to see if the data accord with model expectations (17). Ideal fit values are reported at the bottom of the fit table (Table 3).

Unidimensionality is tested with a principal component analysis (PCA) of the standardized Rasch residuals. A *t*-test was conducted comparing pairs of ability estimates, either loading positively or negatively on the first component of the residuals. The lower limit of the confidence interval for the percentage of significant *t*-tests should be below 5%.

Scale invariance was tested by examining Differential Item Functioning (DIF). PROMs were considered as invariant or free of DIF, if persons with comparable levels of Daily Activities ability (as defined by the two PROMs under consideration in each pairwise comparison) obtained the same score on a given PROM, regardless of group characteristics, e.g. age, gender and country. Should DIF be observed, a comparison was made between unadjusted and adjusted person estimates, the latter derived by splitting items on the group variable (30). In this study, if a paired t-test between the two estimates was significant, a substantive difference was interpreted as an effect size of that difference ≥ 0.1 (31).

A core of six PROMs, referred to as core scale bank, was identified and co-calibrated to define the reference metric. This core scale bank was designed to prevent replicates of PROM's (i.e. the four PROMIS short forms, the two WOMAC-PF forms and the HAQ and MDHAQ), to avoid problems with a breach of the local independence assumption, and so included the WOMAC-RAT, DASH23, PROMIS PF-20, SF36-PF, and WHODAS-PF (with its three domains summated into a single score), and the HAQ (31). The remaining scales were subsequently calibrated onto the metric on an individual basis, calibrating along with the HAQ, anchored to the item parameters of the HAQ from the core set analysis.

Results

Age, gender and disease duration of the sample in each country are given in Table 1. The contribution to the overall sample made by each country for each PROM is shown in Table 2. The raw data are presented in the way that they are traditionally reported, for example, variations of the HAQ are rescored to 0-3, and the SF36-PF to 0-100. Table 3 gives some basic descriptive statistics for each PROM, as well as the SEM and SDD. WOMAC-PF (in either format) and the PROMIS-PF-20 are the most efficient PROMs in that only approximately 11% of the scale would need to be transited to get above the error. In contrast,

the HAQ would need to transit over one fifth of the PROM (21.1%) or the WHODAS-PF one third (33.5 %) to get above the error. In other words, a 15 % score change in the HAQ cannot be statistically detected but would be veiled by measurement error, whereas such a change in the WOMAC-PF can be already detected as statistically significant change.

[Table 1 and Table 2]

Fit of the PROMs to the Rasch model is shown in Table 4. The four PROMIS-PF sets and the SF36-PF had their scores reversed to be consistent with the other PROMs, so that a high score indicates poor functioning. Each row is a pairwise fit of the HAQ plus one other scale, until the final row brings together a number of scales (core set), avoiding putting scales together that are close replicates of one another. All pairs of scales showed fit to the Rasch model, represented by a non-significant CTF, and all pairs were unidimensional. Some DIF was observed and tested to see if this gave rise to significantly different person estimates. Substantive DIF was absent at the pairwise level, for example, the WOMAC-RAT showed a paired t-test significance of 0.83. In the six-PROM core scale bank, the country based DIF for the WOMAC-RAT was still present. Nevertheless, the effect size of the differences (between the unadjusted and adjusted analyses) was just 0.07.

[Table 3 and 4]

Given that all the PROMs tested fit the assumptions of the Rasch model, a transformation table was created. Appendix 1 shows the exchange rates between the eleven PROMs tested (i.e. including the four forms of the PROMIS; the HAQ and MDHAQ; and two forms of the WOMAC-PF), using the interval-scaled Daily Activities Metric as the link. A high score on this Reference metric represents low ability to perform tasks and, conversely, a low score

represents high ability. The HAQ and MDHAQ were scored in their usual way of 0-3, and the four PROMIS short forms and the SF36-PF scores were reversed, so that a high score represents few, if any, limitations in daily activities. For example, a HAQ score of 0.75 is associated with a Reference Metric score of 43.44, as is WOMAC- RAT score of 17, a DASH23 score of 28, and a SF36-PF score of 55. If there is no direct match then the nearest score is taken, for example, a PROMIS-PF-20 of 77, and a WHODAS-PF score of 13. Even where there is no direct match, the link will be accurate within less than one-tenth of a logit. To facilitate access to the Reference metric, Appendix 1 is presented as an Excel supplementary file. Thus, readers can choose to select just those PROMs relevant to their current analysis to obtain the interval-scaled Daily Activities Metric, or compare PROM scores, or both.

Figure 2 shows the operational ranges of the scales in logits, along with the interval-scaled Daily Activities Metric. Most scales measure a similar range, i.e. within ±2 logits, with only slight variations. These variations manifest also in the transformation Table (Appendix 1) where, for example, the SF36-PF has the lowest Reference Metric of all the scales with 14.20 for its score of 100, but only achieves a metric level of 67.45 for its score of zero. Thus, its orientation is slightly to the more able end of the Daily Activities Metric.

[Figure 2]

Discussion

Many of the most widely used PROMs in RA involve the measurement of Activities of Daily Living, sometimes referred to as Physical Function, and are consistent with ICF Chapters d4 Mobility, d5 Self-Care and d6 Domestic Life. In this study, 11 PROMs were shown to map onto a Daily Activities Metric, and each pair of PROMs, with the HAQ as a common PROM

comparator, showed fit to the Rasch model and unidimensionality. A core set of six PROMs also showed such fit. Given the PROMs all measure a similar range on the Daily Activities Metric, then other factors, such as the SEM and SDD can be used to differentiate between PROMs when selecting which to use in clinical practice or research. For example, the selected items of the DASH23 for upper limb therapy and research, the WOMAC-RAT version for lower limb, and the PROMIS-PF-20 for general use would seem to be the better choices among these PROMs. Of particular concern is the SDD of the HAQ, indicating it is less than optimal for detecting a difference compared to other PROMs.

The approach to use just the total scores of the PROMS as items to fit the Rasch model is relatively new (17). Under the Rasch model, sufficiency is explicitly on the total score of the person for the person parameter, and the total score for the item for the item parameter (33). Here the 'item' is a PROM and thus the total score for the PROM (summed over all persons) estimates the scale parameter. Increasingly, studies are published that examine the potential of standardized reporting by linking commonly used questionnaires (34, 35). The present study differs from these studies as the calibration model used here delivers estimates which are independent of the distribution upon which the calibration is based. Such a calibration model requires parameter separation between persons and items (36), which is consistent with applying the Rasch model, as in the current study. Under these circumstances and given the same frame of reference (e.g. health condition group), clinicians and researchers can have confidence that the transformations (by using e.g. a transformation table) apply to their own sample, involving the same frame of reference. Nevertheless, given the availability of different studies linking commonly used questionnaires to enable comparability using IRT and Rasch models, it remains to be investigated to compare the performance of these different approaches.

The limitations of the study arise from a number of technical issues related to the application and interpretation of the results. For example, current software constraints limit the operational range of an item, in the case of RUMM2030, to 100 categories. Thus, the WOMAC-NRS with a range of 170, had to be divided by 1.7 and rounded for fit to the model, and then expanded again for comparability purposes. The use of the transformation Table (Appendix 1) itself is also constrained to where there are complete data, although recent work has shown that if necessary, imputation of missing data (missing completely at random or at random) will not affect the interpretation of fit to the Rasch model (37). Missing data at the scale level is treated in the same way as in item-based analysis, that is, estimates are based on the information available (i.e. the scale is treated as missing for that case), but missing data is always an indicator of the validity of the scale in a given population, irrespective of the analytical strategy chosen. The sample size, while adequate for the Rasch model application, nevertheless is modest compared with other equating studies using different IRT approaches (34, 35), but the latter require much larger sample sizes for their chosen models.

The strengths to the study come from the content comparability checks based on the ICF and the confirmation of unidimensionality of the item sets through the Rasch model. The model itself has sufficiency of the person score, such that the only information required is the total score for the person (33). Thus, clinicians and researchers can simply add up the responses to a set of items and have access to the Daily Activities Metric through the transformation table (Appendix 1). The link to the ICF is also consistent with the latest requirements for e-Health informatics, such that data is recorded based on international standards with the ICF being one of these (15). As such, the approach supports standardized reporting as there is no need to create new PROMs unless there is a sound reason for doing so, for example poor psychometrics in the target population. The scale banking also facilitates comparability of data and results of clinical trials, e.g. through meta-analysis, and patient registries.

Conclusions

Many scales used to assess the impact of RA involve PROMs which ascertain the level of difficulty across a range of everyday activities as described in chapter d4 Mobility, d5 Self-Care and d6 Domestic Life in the ICF. Data from a mix of the most commonly used PROMs in RA have shown that they consistently map onto these chapters. Fit of their data to the Rasch model has shown that in a pairwise fashion, and with a core set of six PROMs, the data satisfied the Rasch model expectations, making their total scores comparable via an interval-scaled Daily Activities Metric. Descriptive analysis of the scales suggested that, given similar operational ranges on the metric, some PROMs displayed much lower SDD's in relation to their operational range, which will have implications for sample size requirements and detection of change.

List of abbreviations

DASH	Disabilities of Arm, Shoulder and Hand
EULAR	European League Against Rheumatism
HADS	Hospital Anxiety and Depression Scale
HAQ	Health Assessment Questionnaire
ICF	International Classification of Functioning, Disability and Health
IQR	Inter-quartile range
LMU	Ludwig-Maximilians-University
MDHAQ	Multidimensional Health Assessment Questionnaire
NEAT	Common-item Non-equivalent Group Design
NRS	numeric rating scale
OA	Osteoarthritis
PCA	Principal Component Analysis
PF	physical function
PROMIS	Patient-Reported Outcomes Measurement Information System
PROMs	Patient Reported Outcome Measure
RA	Rheumatoid Arthritis
RAQoL	Rheumatoid Arthritis Quality of Life Scale
RAT	rating scale
SDD	Smallest Detectable Difference
SEM	Standard Error of Measurement
UK	United Kingdom
WHODAS 2.0	World Health Organization Disability Assessment Schedule Version 2.0
WOMAC	Western Ontario and McMaster Universities Osteoarthritis Index

References

1. Stack RJ, van Tuyl LH, Sloots M, van de Stadt LA, Hoogland W, Maat B, et al. Symptom complexes in patients with seropositive arthralgia and in patients newly diagnosed with rheumatoid arthritis: a qualitative exploration of symptom development. Rheumatol. 2014;53(9):1646-53.
2. Sverker A, Östlund G, Thyberg M, Thyberg I, Valtersson E, Björk M. Dilemmas of participation in everyday life in early rheumatoid arthritis: a qualitative interview study (The Swedish TIRA Project). DisabilRehabil. 2014;37(14):1251-9.
3. Nicassio PM, Kay MA, Custodio MK, Irwin MR, Olmstead R, Weisman MH. An evaluation of a biopsychosocial framework for health-related quality of life and disability in rheumatoid arthritis. J Psychosom Res. 2011;71(2):79-85.
4. Coenen M, Cieza A, Stamm TA, Amann E, Kollerits B, Stucki G. Validation of the International Classification of Functioning, Disability and Health (ICF) Core Set for rheumatoid arthritis from the patient perspective using focus groups. Arthritis Res Ther. 2006;8(4):R84.
5. Gilworth G, Chamberlain MA, Harvey A, Woodhouse A, Smith J, Smyth MG, et al. Development of a work instability scale for rheumatoid arthritis. Arthritis Care Res. 2003;49(3):349-54.
6. Carr A, Hewlett S, Hughes R, Mitchell H, Ryan S, Carr M, et al. Rheumatology outcomes: the patient's perspective. J Rheumatol. 2003;30(4):880-3.
7. Kalyoncu U, Dougados M, Daurès J, Gossec L. Reporting of patient-reported outcomes in recent trials in rheumatoid arthritis: a systematic literature review. Ann Rheumat Dis. 2009;68(2):183-90.
8. Heller J, Shadick NA. Outcomes in rheumatoid arthritis: incorporating the patient perspective. Curr Opin Rheumatol. 2007;19(2):101-5.
9. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum. 1980;23(2):137-45.
10. WHO. International Classification of Functioning, Disability and Health. Geneva: World Health Organization (WHO); 2001.
11. McCollum L, Pincus T. A biopsychosocial model to complement a biomedical model: patient questionnaire data and socioeconomic status usually are more significant than laboratory tests and imaging studies in prognosis of rheumatoid arthritis. Rheumat Dis Clin North Am. 2009;35(4):699-712.
12. Stucki G, Cieza A, Geyh S, Battistella L, Lloyd J, Symmons D, et al. ICF Core Sets for rheumatoid arthritis. J Rehabil Med. 2004;Suppl 44:87-93.
13. Lundgren-Nilsson Å, Dencker A, Palstam A, Person G, Horton MC, Escorpizo R, Küçükdeveci AA, Kutlay S, Elhan AH, Stucki G, Tennant A, Conaghan PG. Patient-reported outcome measures in osteoarthritis: a systematic search and review of their use and psychometric properties. RMD Open 2018;4 (2)
14. Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: time to end malpractice? J Rehabil Med. 2012;44(2):97-8.
15. ISO. Health informatics - Capacity-based eHealth architecture roadmap. Part 2: Architectural components and maturity model. PD ISO/TR 14639-2:2014. UK: International Standard Organisation; 2014.
16. Prodinger B, Tennant A, Stucki G, Cieza A, Üstün TB. Harmonizing routinely collected health information for strengthening quality management in health systems: requirements and practice. J Health Serv Res Policy. 2016; 21(4): 223-8.
17. Andrich D. The Polytomous Rasch Model and the Equating of Two Instruments. In: Christensen KB, Kreiner S, Mesbah M, editors. Rasch Models in Health. London, UK: ILSTE Ltd; 2013. p. 164-96.
18. Hurst N, Jobanputra P, Hunter M, Lambert M, Lochhead A, Brown H, Validity of euroqol—a generic health status instrument—in patients with rheumatoid arthritis. Eeconomic and health outcomes research group, Rheumatol, 1994;33(7):655-62.

19. Zigmond AS, Snaith RP, The hospital anxiety and depression scale, Acta Psychiatr Scand, 1983;67(6):361-70,
20. De Jong Z, Van Der Heijde D, McKenna SP, Whalley D: The Reliability and Construct Validity of The RAQoL: A Rheumatoid Arthritis Specific Quality Of Life Instrument, Brit J Rheumatol 1997; 36: 878-883.
21. Ware JE, Kosinski M, Dewey JE, Gandek B. How to Score Version 2 of the SF-36® Health Survey. Lincoln, RI: Quality Metric Inc.; 2000.
22. Hudak PL, Amadio PC, Bombardier C, Beaton D, Cole D, Davis A, et al. Development of an upper extremity outcome measure: the DASH (Disabilities of the Arm, Shoulder, and Hand). Am J Indust Med. 1996;29(6):602-8.
23. Pincus T, Swearingen C, Wolfe F. Toward a multidimensional health assessment questionnaire (MDHAQ): assessment of advanced activities of daily living and psychological status in the patient-friendly health assessment questionnaire format. Arthritis Rheum 1999;42(10):2220-30.
24. Bellamy N. The WOMAC knee and hip osteoarthritis indices: development, validation, globalization and influence on the development of the AUSCAN hand OA indices. Clin Experim Rheumatol. 2005;23(5):S148.
25. WHO. WHO Disability Assessment Schedule 2.0 (WHODAS2.0). Geneva: World Health Organization; 2013 [cited 2013 Sept. 3rd]; Available from: http://www.who.int/classifications/icf/whodasii/en/index.html.
26. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. J Clin Epi. 2010;63(11):1179-94.
27. Barber et al. (2019) 2019 American College of Rheumatology Recommended Patient-Reported Functional Status Assessment Measures in Rheumatoid Arthritis. Arthritis Care Res, 71(12):1531-1539.
28. Stucki G. et al. ICF Core Sets for rheumatoid arthritis. Disabil Rehabil 2004: 36, S 44; 87-93
29. Masters GN. A Rasch model for partial credit scoring. Psychometrika. 1982;47:149-74.
30. Hagquist C, Andrich D. Recent advances in analysis of differential item functioning in health research using the Rasch model. Health Qual Life Out. 2017;15(1):181.
31. Marais I, Andrich D. Formalizing Dimension and Response Violations of Local Independence in the Unidimensional Rasch Model. J Appl Meas 2008; 9: 200-15
32. Rouquette A, Hardouin J-B, Vanhaesebrouck A, Se´bille V, Coste J (2019). Differential Item Functioning (DIF) in composite health measurement scale: Recommendations for characterizing DIF with meaningful consequences within the Rasch model framework. PLoS ONE 2019; 14 (4): e0215073.
33. Andersen EB. Sufficient statistics and latent trait models. Psychometrika. 1977;42(1):69-81.
34. Cook, KF; Schalet, BD; Kallen, MA; Rutsohn, JP; Cella, D. Establishing a common metric for self-reported pain: linking BPI Pain Interference and SF-36 Bodily Pain Subscale scores to the PROMIS Pain Interference metric. Qual Life Res 2015; 24:2305-2318
35. Oude Vsohaar MAH, Vonkeman HE, Courvoisier D, Finckh A, Gossec L, Leung YY, Michaud K, Pinheiro G, Soriano E, Wulfraat N, Zink A, van de Laar MAFJ. Towards standardized patient reported phyiscal function outcome reporting: linking ten commonly used questionnaires to a common metric. Qual Lif Res 2019; 28(1): 187-197.
36. Andrich D. Rasch models for measurement. Sage University Paper Series on Quantitative Applications in the Social Sciences. California:Sage Publications. 1988.
37. Fellinghauer C, Prodinger B, Tennant A. The Impact of Missing Values and Single Imputation upon Rasch Analysis Outcomes: A Simulation Study. J Appl Meas. 2018;19(1):1-25.

Table Legends

Figure Legends

Appendix

Table 1: Sample characteristics by Country

Country	Mean Age & SD (in years)	% Female	Mean Disease Duration & SD (in years)	Median HAQ	N
Germany	49.0 (13.8)	91.4	13.5 (12.2)	1.0	180
United Kingdom	68.3 (10.0)	74.2	19.8 (13.0)	1.0	535
Turkey	57.5 (11.5)	75.8	13.7 (10.3)	0.8	458
Other Europe	56.9 (12.7)	80.4	11.3 (9.8)	1.5	554
TOTAL		78.4			1727

Table 2: Country Contributions to Scale Bank

Scale	Country	Sample	Median Score	IQR	Difference: Kruskal Wallis p
WHODAS-PF	Turkey	296	12	4-24	-
SF36-PF	UK	368	40	15-65	
	Other Europe	514	35	15-55	0.0283
PROMIS-PF4	Germany	156	15	12-17	
	UK	152	14	9-18	0.3222
PROMIS-PF6	Germany	156	20	16-24	
	UK	152	19	12-35	0.2178
PROMIS-PF10	Germany	156	34	28-41	
	UK	152	32	24-41	0.2788
PROMIS-PF20	Germany	156	72	60-86	
	UK	152	68	50-85	0.1322
DASH23	Germany	155	34	19-53	
	Turkey	115	33	16-56	0.7942
WOMAC-NRS	Germany	153	49	20-90	-
WOMAC-RAT	UK	141	24	7-35	
	Turkey	155	20	11-36	0.9170
MDHAQ	UK	151	1	0.375-1.75	
	Turkey	156	0.63	0.25-1.5	0.0636
HAQ	Germany	176	1	0.5-1.5	
	UK	529	1	0.5-1.75	
	Turkey	457	0.75	0.125-1.5	
	Other Europe	427	1.5	0.875-2.0	0.0001

NRS = Numeric Rating Scale (0-10); RAT = Rating Scale (0-4)

Table 3: Scale precision, ordered by percentage of Smallest Detectable Difference

Scale	Observations	Median	IQR	Min	Max	ά	SEM	Operational range	SDD	% SDD
WOMAC-RAT	296	21.5	9-36	0	85	0.97	2.96	85	8.21	9.66
WOMAC-NRS	153	49.3	20.4-90.1	0	170	0.98	6.77	170	18.76	11.03
PROMIS-PF20	298	70.0	54-86	20	100	0.97	3.19	80	8.83	11.04
DASH23	270	33.0	18-55	0	92	0.97	4.02	92	11.14	12.11
PROMIS-PF10	305	33.0	26-41	10	50	0.95	2.21	40	6.13	15.32
MDHAQ	307	0.9	0.38-1.63	0	3	0.91	0.19	3	0.54	17.82
PROMIS-PF6	308	20.0	14-25	6	30	0.94	1.59	24	4.42	18.42
SF36-PF	882	35.0	15-60	0	100	0.92	7.41	100	20.53	20.53
HAQ	1589	1.0	0.5-1.75	0	3	0.92	0.23	3	0.63	21.08
PROMIS-PF4	308	14.0	10-18	4	20	0.92	1.33	16	3.68	23.02
WHODAS-PF	296	12.0	4-24	0	52	0.94	6.28	52	17.40	33.47

IQR=Inter-quartile range; ά = Cronbach's alpha; SEM = Standard Error of Measurement; %SDD = % of operational range that is the Smallest Detectable Difference.

Table 4. Fit of scales to the Rasch model

#	Fit of the HAQ	Reliability	Conditional Test of Fit p value	Unidimensionality t-test % at > 0.05 (LCI)	DIF Present	Substantive DIF	N
1							
2	WOMAC-RAT	0.80	0.4504	0.69	Age & Country	-	290
3	WOMAC-NRS	0.93	0.1408	1.36	-		150
4	DASH23	0.96	0.9910	3.08	Gender	-	266
5	PROMIS20	0.98	0.9999	5.44 (3.7)	-		294
6	PROMIS10	0.96	0.9933	4.01	-		299
7	PROMIS6	0.93	0.8095	3.36	-		304
8	PROMIS4	0.91	0.8601	3.10	-		290
9	SF36 PF	0.88	0.5783	1.96	country	-	776
10	WHODAS	0.75	0.9933	0.37	Age & Gender	-	295
	1,2,4,5,9,10	0.87	0.1218	-	Country	-	1665
	Ideal Values	**>0.7**	**>0.05**	**<5.0 (LCI <5.0)**	**Absent**	**Absent**	

LCI = Lower Confidence Interval; DIF = Differential item functioning

Figure 1: Overview data structure

Light grey represents previously collected data used in this study for secondary analysis, dark grey reflects data that has been newly collected specifically for this project.

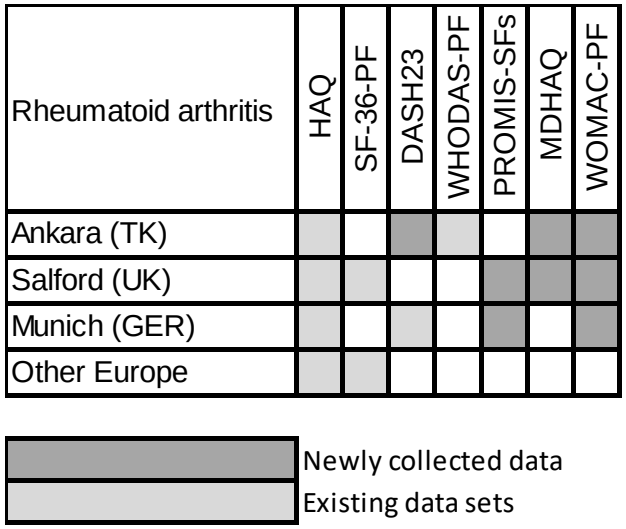
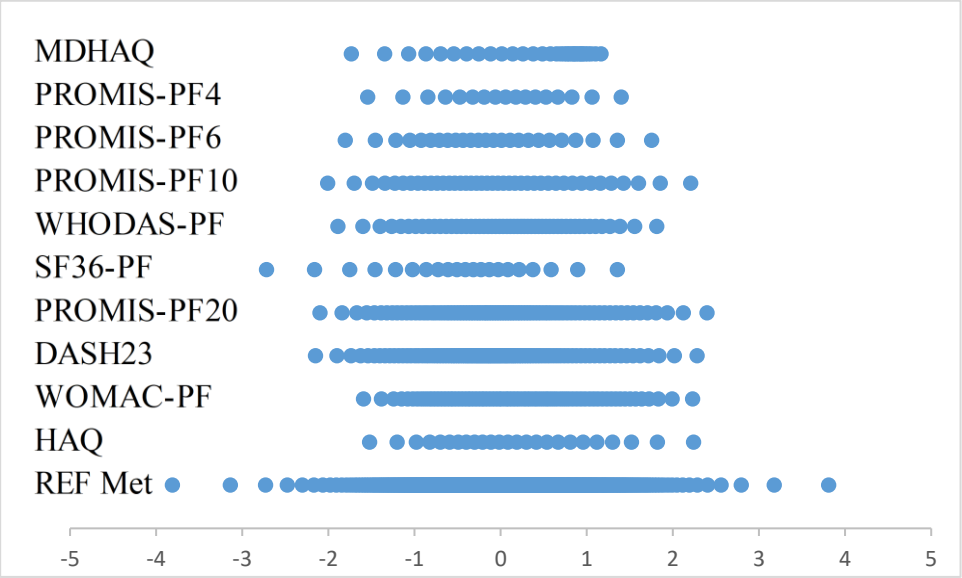


Figure 2: Operational Widths of Scales on the Interval-Scaled Daily Activities Reference Metric



Key: REF Met = Reference Metric

Appendix 1: Scale-to-scale transformation table with the Daily Activities Reference Metric

REFMET	HAQ	MDHAQ	WOMAC-RAT	WOMAC-NRS	DASH23	PROMIS20	PROMIS10	PROMIS6	PROMIS4	SF36 PF	WHODAS
0.00											
8.81											
14.20										100	
17.51											
19.82											
21.55					0	100				95	
22.93											
24.06							50				
25.00					1						
25.81						99					0
26.54								30			
27.17		0			2					90	
27.74						98	49				
28.25											
28.73				0	3						
29.17			0								1
29.58						97			20		
29.96	0				4						
30.33											
30.67					5	96	48				
31.00								29		85	
31.31				2							
31.61					6						2
31.89			1			95					
32.17											
32.44		0.1			7		47		19		
32.69						94					

32.94												
33.19				3	8							
33.43						93						3
33.65			2									
33.87					9	92	46			80		
34.09								28				
34.30	0.125			5								
34.51					10							
34.72												
34.92			3	7		91						4
35.12					11		45					
35.30												
35.50												
35.68						90						
35.87			4	9	12							
36.04		0.2										5
36.22								27				
36.39					13	89	44					
36.56			5	10						75		
36.72												
36.89												
37.05				12	14	88						6
37.20	0.25											
37.36			6				43					
37.52					15							
37.68				14		87						
37.83								26				
37.98			7									
38.12				15	16							7
38.27						86						

38.41							42					
38.56			8		17							
38.70		0.3		17							70	
38.85												
38.98						85				18		
39.12			9	19	18							8
39.25	0.375											
39.38							41	25				
39.53				20	19	84						
39.66			10									
39.79												
39.92				22								9
40.04												
40.17			11		20	83						
40.30												
40.43				24			40					
40.55					21						65	
40.68			12			82		24				
40.80	0.5			26								10
40.92		0.4										
41.05					22							
41.17				27		81						
41.29			13				39					
41.40					23							
41.52												
41.64				29						17		11
41.76			14			80						
41.88					24							
41.99				31				23			60	
42.10												

42.22	0.625						79	38					
42.34			15	32	25								
42.44													
42.56													12
42.66					26								
42.78				34		78							
42.89		0.5	16										
43.01								37					
43.11				36	27				22				
43.22						77							
43.32													13
43.44	0.75		17		28							55	
43.54				37									
43.65						76							
43.75											16		
43.86				39	29			36					
43.96			18										
44.07													14
44.17				41	30	75							
44.28									21				
44.37													
44.48			19										
44.58				43	31	74							
44.69								35				50	
44.78	0.875	0.6											15
44.88				44	32								
44.99			20										
45.08						73							
45.18				46	33								
45.28													

45.38										
45.47			21	48	34	72	34	20		16
45.58										
45.67										
45.77				49					15	
45.87					35	71			45	
45.96	1									
46.06			22	51						
46.15					36					17
46.25										
46.34				53		70	33			
46.44					37					
46.54			23							
46.63		0.7		54				19		
46.72					38	69				
46.82										18
46.92				56						
47.01					39				40	
47.10			24			68	32			
47.19				58						
47.28	1.125				40					
47.38										
47.47				60		67				19
47.56			25						14	
47.65				61	41					
47.74								18		
47.85						66				
47.94				63	42		31			
48.03										
48.12			26		43					20

48.22				65							
48.31						65				35	
48.40				66							
48.49	1.25	0.8			44						
48.58											
48.67			27	68		64					21
48.77					45		30				
48.86				70							
48.95						63		17			
49.04					46						
49.13			28	71							
49.23									13		
49.32				73	47						22
49.41						62					
49.50				75							
49.59					48		29				
49.69			29							30	
49.78	1.375			77		61					
49.87											23
49.96				78	49						
50.05											
50.14				80				16			
50.24		0.9	30		50	60					
50.33				82							
50.42							28				24
50.53				83	51						
50.62						59					
50.71			31	85							
50.80					52				12		
50.89				87							

50.98						58					25
51.09	1.5			88						25	
51.18					53						
51.27			32	90			27				
51.36						57					
51.47				92	54			15			26
51.56											
51.65				94							
51.76					55	56					
51.85		1	33	95							
51.96											
52.05				97							27
52.15					56	55	26				
52.24				99							
52.35			34	100					11		
52.44					57						
52.55	1.625			102		54					28
52.64											
52.74				104						20	
52.83				105	58			14			
52.94			35			53					
53.04				107							
53.15					59		25				29
53.24				109							
53.35											
53.45		1.1	36	111		52					
53.56				112	60						
53.66											30
53.77				114					10		
53.87						51					

53.98	1.75		37	116	61							
54.08							24					
54.19				117								31
54.29				119	62	50		13				
54.40												
54.50			38	121								
54.61												
54.72				122	63	49						32
54.83				124								
54.93		1.2									15	
55.05			39	126			23					
55.16					64							
55.28				128		48						33
55.38				129						9		
55.50	1.875											
55.60				131	65							
55.72			40			47						
55.84				133				12				34
55.95												
56.06				134	66							
56.18				136		46	22					
56.30			41									35
56.42		1.3		138								
56.54					67							
56.65				139								
56.77			42	141		45						
56.89										8		36
57.01				143	68							
57.14	2											
57.26				145		44	21					

57.38			43								
57.51				146	69			11			37
57.62		1.4									
57.76				148		43				10	
57.87				150							
58.01			44		70						
58.14											38
58.25				151							
58.39						42					
58.52		1.5	45	153	71		20				
58.65											
58.78	2.125			155					7		39
58.91						41					
59.04				156							
59.17		1.6	46		72						
59.32								10			40
59.45				158							
59.58						40					
59.72		1.7	47	160	73		19				
59.86											
60.00											41
60.14		1.8		162		39					
60.28					74						
60.42			48								
60.56		1.9									
60.71	2.25			163		38					42
60.85		2							6		
61.00			49		75		18				
61.16		2.1									
61.30				165							43

61.44		2.2					37		9			
61.60			50		76							
61.75											5	
61.90		2.3										
62.06				167			36					44
62.22		2.4	51		77							
62.38								17				
62.53	2.375	2.5										
62.69												
62.85		2.6	52		78		35					45
63.01												
63.18												
63.33		2.7		168								
63.50			53				34					
63.67					79			16				46
63.85		2.8										
64.02									8	5		
64.19			54									
64.36					80		33					
64.54		2.9										47
64.71	2.5											
64.90			55									
65.08					81		32					
65.26		3						15				
65.45			56	170								48
65.64												
65.83					82		31					
66.02												
66.22			57									
66.42					83							

66.61						30					49
66.82			58				14				
67.02	2.625										
67.23											
67.45			59		84	29				0	
67.66											
67.89								7			
68.11											50
68.33			60		85	28			4		
68.57											
68.81							13				
69.04			61								
69.29					86	27					
69.54	2.75										
69.80			62								
70.08					87						
70.35						26					51
70.63			63								
70.93											
71.23					88	25	12				
71.56			64								
71.89											
72.24											
72.61			65		89	24					
73.01								6			
73.43											
73.88	2.875		66		90	23					52
74.37							11				
74.91											
75.50						22					

76.14			67		91						
76.89											
77.76						21					
78.77	3		68				10				
80.01					92						
81.56						20					
83.62											
86.68											
91.71											
100.00											